What is Twitter, a Social Network or a News Media?

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon
Department of Computer Science, KAIST
335 Gwahangno, Yuseong-gu, Daejeon, Korea
{haewoon, chlee, hosung}@an.kaist.ac.kr, sbmoon@kaist.edu

ABSTRACT

Twitter, a microblogging service less than three years old, commands more than 41 million users as of July 2009 and is growing fast. Twitter users tweet about any topic within the 140-character limit and follow others to receive their tweets. The goal of this paper is to study the topological characteristics of Twitter and its power as a new medium of information sharing.

We have crawled the entire Twitter site and obtained 41.7 million user profiles, 1.47 billion social relations, $4,262$ trending topics, and 106 million tweets. In its follower-following topology analysis we have found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks [28]. In order to identify influentials on Twitter, we have ranked users by the number of followers and by PageRank and found two rankings to be similar. Ranking by retweets differs from the previous two rankings, indicating a gap in influence inferred from the number of followers and that from the popularity of one's tweets. We have analyzed the tweets of top trending topics and reported on their temporal behavior and user participation. We have classified the trending topics based on the active period and the tweets and show that the majority (over 85%) of topics are headline news or persistent news in nature. A closer look at retweets reveals that any retweeted tweet is to reach an average of $1,000$ users no matter what the number of followers is of the original tweet. Once retweeted, a tweet gets retweeted almost instantly on next hops, signifying fast diffusion of information after the 1st retweet.

To the best of our knowledge this work is the first quantitative study on the entire Twittersphere and information diffusion on it.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Human Factors, Measurement

Keywords

Twitter, Online social network, Reciprocity, Homophily, Degree of separation, Retweet, Information diffusion, Influential, PageRank

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

1. INTRODUCTION

Twitter, a microblogging service, has emerged as a new medium in spotlight through recent happenings, such as an American student jailed in Egypt and the US Airways plane crash on the Hudson river. Twitter users follow others or are followed. Unlike on most online social networking sites, such as Facebook or MySpace, the relationship of following and being followed requires no reciprocation. A user can follow any other user, and the user being followed need not follow back. Being a follower on Twitter means that the user receives all the messages (called *tweets*) from those the user follows. Common practice of responding to a tweet has evolved into well-defined markup culture: RT stands for retweet, '@' followed by a user identifier address the user, and '#' followed by a word represents a hashtag. This well-defined markup vocabulary combined with a strict limit of 140 characters per posting conveniences users with brevity in expression. The *retweet* mechanism empowers users to spread information of their choice beyond the reach of the original tweet's followers.

How are people connected on Twitter? Who are the most influential people? What do people talk about? How does information diffuse via retweet? The goal of this work is to study the topological characteristics of Twitter and its power as a new medium of information sharing. We have crawled 41.7 million user profiles, 1.47 billion social relations, and 106 million tweets[1]. We begin with the network analysis and study the distributions of followers and followings, the relation between followers and tweets, reciprocity, degrees of separation, and homophily. Next we rank users by the number of followers, PageRank, and the number of retweets and present quantitative comparison among them. The ranking by retweets pushes those with fewer than a million followers on top of those with more than a million followers. Through our trending topic analysis we show what categories trending topics are classified into, how long they last, and how many users participate. Finally, we study the information diffusion by retweet. We construct retweet trees and examine their temporal and spatial characteristics. To the best of our knowledge this work is the first quantitative study on the entire Twittersphere and information diffusion on it.

This paper is organized as follows. Section 2 describes our data crawling methodology on Twitter's user profile, trending topics, and tweet messages. We conduct basic topological analysis of the Twitter network in Section 3. In Section 4 we apply the PageRank algorithm on the Twitter network and compare its outcome against ranking by retweets. In Section 5 we study how their popularity rises and falls among users over time. In Section 6 we focus information diffusion through retweet trees. Section 7 covers related work and puts our work in perspective. In Section 8 we conclude.

[1]We make our dataset publicly available online at:
http://an.kaist.ac.kr/traces/WWW2010.html

2. TWITTER SPACE CRAWL

Twitter offers an Application Programming Interface (API) that is easy to crawl and collect data. We crawled and collected profiles of all users on Twitter starting on June 6th and lasting until June 31st, 2009. Additionally, we collected profiles of users who mentioned trending topics until September 24th, 2009. On top of user profiles we also collected popular topics on Twitter and tweets related to them. Below we describe in detail how we collected user profiles, popular topics, and related tweets.

2.1 Data Collection

User Profile

A Twitter user keeps a brief profile about oneself. The public profile includes the full name, the location, a web page, a short biography, and the number of tweets of the user. The people who follow the user and those that the user follows are also listed. In order to collect user profiles, we began with Perez Hilton who has over one million followers and crawled breadth-first along the direction of followers and followings. Twitter rate-limits $20,000$ requests per hour per whitelisted IP. Using 20 machines with different IPs and self-regulating collection rate at $10,000$ requests per hour, we collected user profiles from July 6th to July 31st, 2009. To crawl users not connected to the Giant Connected Component of the Twitter network, we additionally collected profiles of those who refer to trending topics in their tweets from June to August. The final tally of user profiles we collected is 41.7 million. There exist 1.47 billion directed relations of following and being followed.

Trending Topics

Twitter tracks phrases, words, and hashtags that are most often mentioned and posts them under the title of "trending topics" regularly. A hashtag is a convention among Twitter users to create and follow a thread of discussion by prefixing a word with a '#' character. The social bookmarking site Del.icio.us also uses the same hashtag convention.

Twitter shows a list of top ten trending topics of the moment on a right sidebar on every user's homepage by default, unless set otherwise. Twitter does not group similar trending topics and, when Michael Jackson died, most of the top ten trending topics were about him: Michael Jackson, MJ, King of Pop, etc. Although the exact mechanism of how Twitter mines the top ten trending topics is not known, we believe the trending topics are a good representation, if not complete, of issues that draw most attention and have decided to crawl them. We collected the top ten trending topics every five minutes via Twitter Search API [36]. The API returns the trending topic title, a query string, and the time of the API request. We used the query string to grab all the tweets that mention the trending topic. In total we have collected $4,262$ unique trending topics and their tweets.

Once any phrase, word, or hashtag appears as a top trending topic, we follow it for seven more days after it is taken off the top ten trending topics' list.

Tweets

On top of trending topics, we collected all the tweets that mentioned the trending topics. The Twitter Search API returns a maximum number of $1,500$ tweets per query. We downloaded the tweets of a trending topic at every 5 minute interval. That is, we captured at most 5 tweets per second. We collected the full text, the author, the written time, the ISO standard language code of a tweet, as well as the receiver, if the tweet is a reply, and the third party application, such as Tweetie.

2.2 Removing Spam Tweets

Spam tweets have increased in Twitter as the popularity of Twitter grows as reported in [35]. As spam web page farms undermine the accuracy of PageRank and spam keywords inserted in web pages hinder relevant web page extraction, spam tweets add noise and bias in our analysis. The Twitter Support Team suspends any user reported to be a spammer. Still unreported spam tweets can creep into our data. In order to remove spam tweets, we employ the well-known mechanism of the FireFox add-on, Clean Tweets [6]. Clean Tweets filters tweets from users who have been on Twitter for less than a day when presenting Twitter search results to FireFox. It also removes those tweets that contain three or more trending topics. We use the same mechanisms in removing spam tweets from our data.

Before we set the threshold of the trending topics to 3 in our spam filtering, we vary the number from 3 to 10 and see the change in the number of identified spam tweets. As we decrease the threshold from 10 to 8, 5, and 3, an order of magnitude more tweets are categorized as spam each time and removed. A tweet is limited to 140 characters and most references to other web pages are abbreviated via URL shortening services (e.g., http://www.tiny.cc/ and http://bit.ly) so that readers could not guess where the references point at. This is an appealing feature to spammers and spammers add as many trending topics as possible to appear in top results for any search in Twitter. There are $20,217,061$ tweets with more than 3 trending topics and $1,966,461$ unique users are responsible for those tweets. For the rest of the paper we remove those tweets from collected tweets. The final number of collected tweets is 106 millions.

3. ON TWITTERERS' TRAIL

We begin our analysis of Twitter space with the following question: How the directed relationship in Twitter impacts the topological characteristics? Numerous social networks have been analyzed and compared against each other. Before we delve into the eccentricities and peculiarities of Twitter, we run a batch of well-known analysis and present the summary.

3.1 Basic Analysis

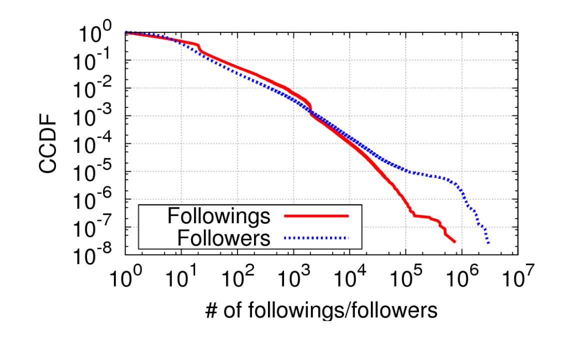


Figure 1: Number of followings and followers

We construct a directed network based on the following and followed and analyze its basic characteristics. Figure 1 displays the distribution of the number of followings as the solid line and that of followers as the dotted line. The y-axis represents complementary cumulative distribution function (CCDF). We first explain the distribution of the number of followings. There are noticeable glitches in the solid line. The first occurs at $x = 20$. Twitter recommends

an initial set of 20 people a newcomer can follow by a single click and quite a few people take up on the offer. The second glitch is at around $x = 2000$. Before 2009 there was an upper limit on the number of people a user could follow [12]. Twitter removed this cap and there is no limit now. The glitch represents the gap in the momentum of network building inflicted by the upper limit. A very small number of users follow more than $10,000$. They are mostly official pages of politicians and celebrities who need to offer some form of customer service.

The dashed line in Figure 1 up to $x = 10^5$ fits to a power-law distribution with the exponent of 2.276. Most real networks including social networks have a power-law exponent between 2 and 3. The data points beyond $x = 10^5$ represent users who have many more followers than the power-law distribution predicts. Similar tail behavior in degree distribution has been reported from Cyworld in [1] but not from other social networks. The common characteristics between Twitter and Cyworld are that many celebrities are present and they readily form online relations with their fans.

There are only 40 users with more than a million followers and all of them are either celebrities (e.g. Ashton Kutcher, Britney Spears) or mass media (e.g. the Ellen DeGeneres Show, CNN Breaking News, the New York Times, the Onion, NPR Politics, TIME). The top 20 are listed in Figure 7. Some of them follow their followers, but most of them do not (the median number of followings of the top 40 users is 114, three orders of magnitude smaller than the number of followers). We revisit the issue of reciprocity in Section 3.3.

3.2 Followers vs. Tweets

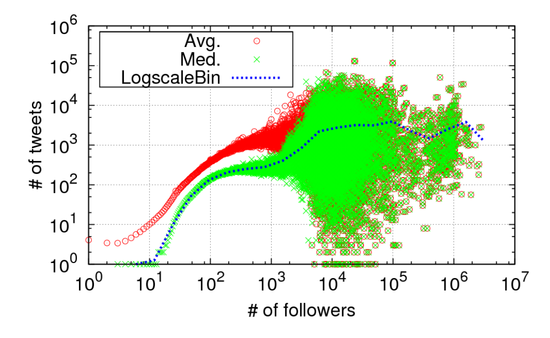


Figure 2: The number of followers and that of tweets per user

In order to gauge the correlation between the number of followers and that of written tweets, we plot the number of tweets (y) against the number of followers a user has (x) in Figure 2. We bin the number of followers in logscale and plot the median per bin in the dashed line. The majority of users who have fewer than 10 followers never tweeted or did just once and thus the median stays at 1. The average number of tweets against the number of followers per user is always above the median, indicating that there are outliers who tweet far more than expected from the number of followers. The median number of tweets stays relatively flat in $x = 100$ to $1,000$, and grows by an order of magnitude for $x > 5,000$.

We gauge the inclination to be active by the number of people a user follows and plots in Figure 3. As pointed out in Figure 1 irregularities at $x = 20$ and $x = 2000$ are observed. Yet the graph plunges at a few more points, $x = 250, 500, 2000, 5000$. We conjecture that they are spam accounts, as many of them have disappeared as of October 2009. We also bin the number of followers in logscale and plot the median per bin in the dashed line. The dashed

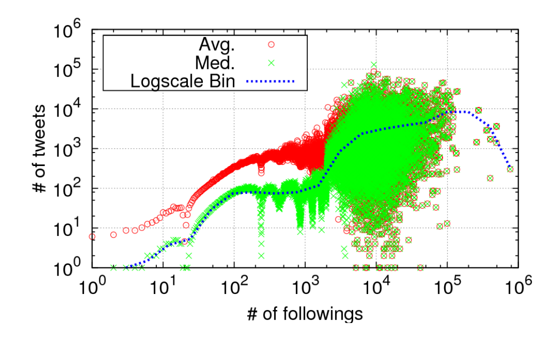


Figure 3: The number of followings and that of tweets per user

line shows a positive trend, while the line is flat between 100 and $1,000$. As in Figure 2 the number of tweets increases by an order of magnitude as the number of followings goes over $5,000$.

Figures 2 and 3 demonstrate that the median number of tweets increases up to $x = 10$ against both the numbers of followers and followings and remains relatively flat up till $x = 100$. Then beyond $x = 5,000$ the number of tweets increases by an order of magnitude or more. Our numbers do not state causation of the peer pressure, but only state the correlation between the numbers of tweets and followers.

3.3 Reciprocity

In Section 3.1 we briefly mention that top users by the number of followers in Twitter are mostly celebrities and mass media and most of them do not follow their followers back. In fact Twitter shows a low level of reciprocity; 77.9% of user pairs with any link between them are connected one-way, and only 22.1% have reciprocal relationship between them. We call those *r-friends* of a user as they reciprocate a user's following. Previous studies have reported much higher reciprocity on other social networking services: 68% on Flickr [4] and 84% on Yahoo! 360 [18].

Moreover, 67.6% of users are not followed by any of their followings in Twitter. We conjecture that for these users Twitter is rather a source of information than a social networking site. Further validation is out of the scope of this paper and we leave it for future work.

3.4 Degree of Separation

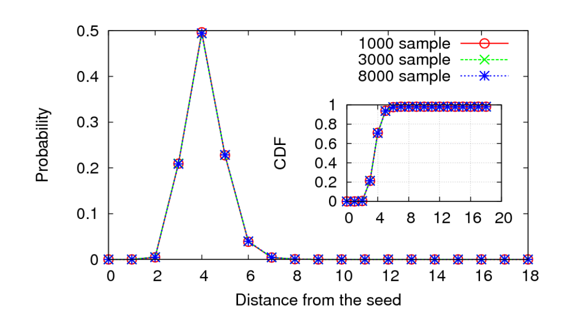


Figure 4: Degree of separation

The concept of degrees of separation has become a key to understanding the societal structure, ever since Stanley Milgram's famous 'six degrees of separation' experiment [27]. In his work he reports that any two people could be connected on average within

six hops from each other. Watts and Strogatz have found that many social and technological networks have small path lengths [37] and call them a 'small-world'. Recently, Leskovec and Horvitz report on the MSN messenger network of 180 million users that the median and the 90% degrees of separation are 6. and 7.8, respectively[22].

The main difference between the above networks and Twitter is the directed nature of Twitter relationship. In MSN a link represents a mutual agreement of a relationship, while on Twitter a user is not obligated to reciprocate followers by following them. Thus a path from a user to another may follow different hops or not exist in the reverse direction.

As only 22.1% of user pairs are reciprocal, we expect the average path length between two users in Twitter to be longer than other known networks. To estimate the path-length distribution we use the same random sampling approach as in [1]. We choose a seed at random and obtain the distribution of shortest paths between the seed and the rest of the network by breadth-first search. Figure 4 exhibits the distributions of the shortest paths in Twitter with $1,000$, $3,000$ and $8,000$ seeds. All three distributions overlap almost completely, showing that the sample size of $8,000$ is large enough. The median and the mode of the distribution are both 4, and the average path length is 4.12. The 90th percentile distance, known as the effective diameter [23], is 4.8. For 70.5% of node pairs, the path length is 4 or shorter, and for 97.6% it is 6 or shorter. There are 1.8% users who have no incoming edge, and the longest path in our samples is 18.

The average path length of 4.12 is quite short for the network of Twitter size, and is the opposite of our expectation on a directed graph. This is an interesting phenomenon that may bespeak for the Twitter's role other than social networking. People follow others not only for social networking, but for information, as the act of following represents the desire to receives all tweets by the person. We note that information is to flow over less than 5 or fewer hops between 93.5% of user pairs, if it is to, taking fewer hops than on other known social networks.

3.5 Homophily

Homophily is a tendency that "a contact between similar people occurs at a higher rate than among dissimilar people" [26]. Weng *et al.* have reported that two users who follow reciprocally share topical interests by mining their 50 thousands links [38]. Here we investigate homophily in two contexts: geographic location and popularity. Twitter users self-report their location. It is hard to parse location due to its free form. Instead, we consider the time zone of a user as an approximate indicator for the location of the user. A user chooses one of the 24 time zones around the world [2]. We drop those users without time zone information in this evaluation. We calculate the time differences between a user and r-friends and compute the average. We plot the median time different versus the number of r-friends in Figure 5.

We observe that the median time difference between a user and r-friends slowly increases as the number of r-friends increases and disperses beyond $x = 2,000$. For those users with $2,000$ r-friends or fewer, the median time differences of the user and r-friends stays below 3 hours. For those with 50 or fewer r-friends, the mean time difference is only about 1.07 hours. For 75% of users the time difference is 3.00 hours or less. For some users who have more than $5,000$ r-friends, the average time difference is more than 6 hours.

[2]We are aware of a campaign to urge users to alter their time zones during the Iranian election in June 2009 [31]. However, we have no means to verify the true time zone of a user and use our data as is.

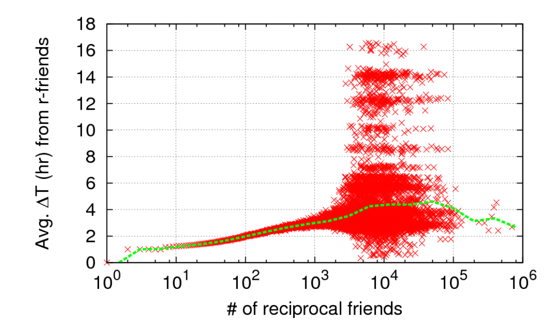


Figure 5: The average time differences between a user and r-friends

This can be interpreted as a large following in another continent. We conclude that Twitter users who have reciprocal relations of fewer than $2,000$ are likely to be geographically close.

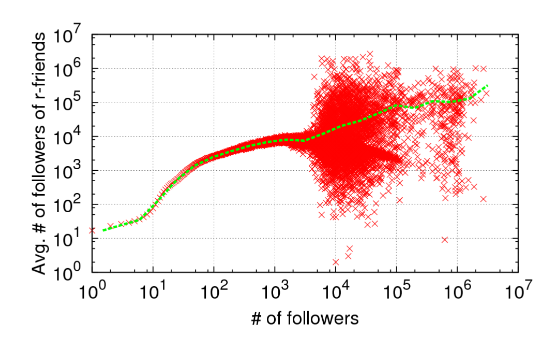


Figure 6: The average number of followers of r-friends per user

Next, we consider the number of followers of a user as an indicator of the user's popularity. Then we ask "Does a user of certain popularity follow other users of similar popularity and they reciprocate?" This question is similar to degree correlation. The degree correlation compares a node's degree against those of its neighbors, and tells whether a hub is likely to connect other hubs rather than low-degree nodes in an undirected network. The positive trend in degree correlation is called assortativity and is known as one of the characteristic features of human social networks [28]. However, it is feasible only in undirected graphs and does not apply to Twitter.

Figure 6 plots the mean of average numbers of followers of r-friends against the number of followers. We see positive correlation slightly below $x = 1,000$ and dispersion beyond that point.

In this section we have looked into homophily from two perspectives: geographic location and the number of r-friends' followers. We observe that users with followers $1,000$ or less are likely to be geographically close to their r-friends and also have similar popularity with their r-friends. Here we have not included the unreciprocated directed links and focused on r-friends. In a way we looked at the social networking aspect of Twitter and found some level of homophily.

In summary Twitter diverges from well-known traits of social networks: its distribution of followers is not power-law, the degree of separation is shorter than expected, and most links are not reciprocated. But if we look at reciprocated relationships, then they exhibit some level of homophily.

Rank	Ranking by # of followers			Ranking by PageRank in the following/follower network			Ranking by # of retweet in the diffusion network		
	ID	Name	Remark	ID	Name	Remark	ID	Name	Remark
1	aplusk	ashton kutcher	actor	aplusk	ashton kutcher	actor	mashable	Pete Cashmore	news on social media
2	britneyspears	Britney Spears	musician	BarackObama	Barack Obama	president of U.S.	BreakingNews	BNO News	news
3	TheEllenShow	Ellen DeGeneres	show host	cnnbrk	CNN Breaking News	news	tweetmeme	TweetMeme	news on Twitter
4	cnnbrk	CNN Breaking News	news	TheEllenShow	Ellen DeGeneres	show host	oxfordgirl	oxfordgirl	journalist
5	Oprah	Oprah Winfrey	show host	britneyspears	Britney Spears	musician	cnnbrk	CNN Breaking News	news
6	twitter	Twitter	subject of this paper	Oprah	Oprah Winfrey	show host	TechCrunch	Michael Arrington	news on technology
7	BarackObama	Barack Obama	president of U.S.	THE_REAL_SHAQ	THE_REAL_SHAQ	sports star	myfabolouslife	Fabolous	musician
8	RyanSeacrest	Ryan Seacrest	show host	johncmayer	John Mayer	musician	nytimes	The New York Times	news
9	THE_REAL_SHAQ	THE_REAL_SHAQ	sports star	twitter	Twitter	subject of this paper	lilduval	lil duval	comedian
10	KimKardashian	Kim Kardashian	model	RyanSeacrest	Ryan Seacrest	show host	IranRiggedElect	Iran	about Iran
11	johncmayer	John Mayer	musician	lancearmstrong	Lance Armstrong	sports star	espn	ESPN Sports News	news
12	mrskutcher	Demi Moore	actress	jimmyfallon	Jimmy Fallon	actor	persiankiwi	persiankiwi	about Iran
13	iamdiddy	iamdiddy	musician	iamdiddy	iamdiddy	musician	aplusk	ashton kutcher	actor
14	jimmyfallon	Jimmy Fallon	actor	mrskutcher	Demi Moore	actress	StopAhmadi	Raymond Jahan	about Iran
15	lancearmstrong	Lance Armstrong	sports star	PerezHilton	Perez Hilton	power blogger	Alyssa_Milano	Alyssa Milano	actress
16	algore	Al Gore	politician	nytimes	The New York Times	news	huffingtonpost	HuffingtonPost.com	news
17	mileycyrus	Miley Cyrus	actress / musician	mileycyrus	Miley Cyrus	actress / musician	iamdiddy	iamdiddy	musician
18	nytimes	The New York Times	news	stephenfry	Stephen Fry	actor	iranbaan	Fershteh Ghazi	about Iran
19	coldplay	Coldplay	musician	TheOnion	The Onion	news	nprnews	NPR News	news
20	TheOnion	The Onion	news	KimKardashian	Kim Kardashian	model	PerezHilton	Perez Hilton	power blogger

Figure 7: Top 20 users ranked by the number of followers, PageRank in the follower network, and the number of retweets

4. RANKING TWITTER USERS

The popularity of a Twitter user can be easily estimated by the number of followers. The top 20 users by the number of followers are listed in Figure 7. We call them List #1. All are either celebrities (actors, musicians, politicians, show hosts, and sports stars) or news media. However, the number of followers alone does not reflect the influence a user exerts when the user's tweet is retweeted many times or is simply followed by other influential people: it is not a comprehensive measure. This problem of ranking nodes based on the topological dependence in a network is similar to ranking web pages based on its connectivity. Google uses the PageRank algorithm to rank web pages in their search results [29]. The key idea behind PageRank is to allow propagation of influence along the network of web pages, instead of just counting the number of other web pages pointing at the web page. In this section we rank users by the PageRank algorithm and also by the number of retweets and compare the outcome.

4.1 By PageRank

We first apply PageRank to the network of followings and followers. In this network a node maps to a user, and every directed edge maps to a user following another. Top 20 ranked users are shown in Figure 7. Let us name this List #2. This top 20 list has the same users as List #1 except for Perez Hilton and Stephen Fry. Al Gore and The Onion are dropped from List #1 and some have changed ranks. Although the two lists do not match exactly, users are ranked similarly by the number of followers and PageRank.

4.2 By the Retweets

The number of retweets for a certain tweet is a measure of the tweet's popularity and in turn of the tweet writer's popularity. Here we rank users by the total number of retweets. The rightmost column in Figure 7 lists the top 20 users by the number of retweets. Only 4 out of 20 users are common in all three rankings. The ranking by the retweets only has one additional user (Perez Hilton) that is common with the PageRank list. The rest are not in either of the first two rankings. A closer look at the users reveals that 4 users rose to fame due to active tweeting during and after the Iran election on June 12th, 2009. There are mainstream news media that rise in ranking by the retweets: The Breaking News Wire, ESPN Sports News, the Huffington Post, and NPR News. It is hard to interpret their rise in retweet ranking, but their rise speaks that followers of these media think that tweets of these media are worth propagating. Quality, timeliness, and coverage of reporting are all candidate factors that we leave for future investigation. A few users, oxfordgirl, Pete Cashmore, and Michael Arrington, can be categorized as independent news media based on online distribution. Ranking by the retweets shows the rise of alternative media in Twitter.

4.3 Comparison among Rankings

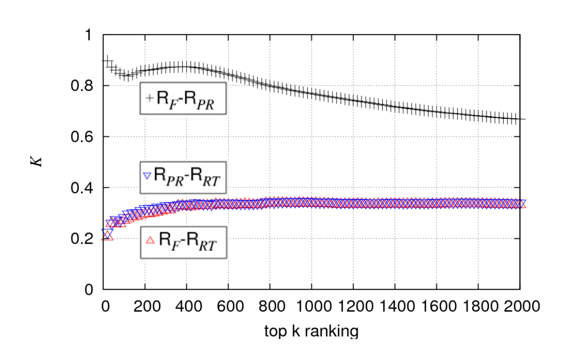


Figure 8: Comparison among rankings

In this section we present a quantitative comparison between the three rankings. We compare the three rankings by the number of followers ($\mathbf{R}_F$), PageRank ($\mathbf{R}_{PR}$) and the number of retweets ($\mathbf{R}_{RT}$) in terms of Fagin *et al.*'s generalized Kendall's tau [8]. Kendall's tau is a measure of rank correlation [16], but original Kendall's tau has the limitation that rankings in consideration must have the same elements. Fagin *et al.* overcome the limitation by comparing only top k lists and adding a penalty parameter, p. We use the "optimistic approach" of Kendall's tau $K_\tau^{(p)}$ with penalty $p = 0$ considering two rankings as $\mathbf{R}_1$ and $\mathbf{R}_2$.

$$K_\tau^{(0)}(\mathbf{R}_1, \mathbf{R}_2) = \sum_{r1, r2 \in \mathbf{R}_1 \cup \mathbf{R}_2} \bar{K}_{r_1, r_2}(\mathbf{R}_1, \mathbf{R}_2) \quad (1)$$

where $\bar{K}_{r_1,r_2}(\mathbf{R}_1, \mathbf{R}_2) = 1$, if (i) r_1 is only in one list and r_2 is in the other list; (ii) r_1 is ranked higher than r_2 in one list and only r_2 appears in the other list; or (iii) r_1 and r_2 are in both lists but in the opposite order. Otherwise, $\bar{K}_{r_1,r_2}(\mathbf{R}_1, \mathbf{R}_2) = 0$. We use the

normalized distance, K, computed as below [25].

$$K = 1 - \frac{K_{\tau}^{(0)}(\mathbf{R}_1, \mathbf{R}_2)}{k^2} \qquad (2)$$

where k is the number of elements in each ranking. The range of K is from 0 to 1. $K = 0$ means complete disagreement, and $K = 1$ means complete agreement.

We plot K for three pairs of rankings varying k from 20 to $2,000$ in Figure 8. We note that $\mathbf{R}_F$-$\mathbf{R}_{PR}$ pair has high K over 0.6 but both $\mathbf{R}_F$-$\mathbf{R}_{RT}$ and $\mathbf{R}_{PR}$-$\mathbf{R}_{RT}$ pairs have low K under 0.4. This means that $\mathbf{R}_F$ and $\mathbf{R}_{PR}$ are similar, but $\mathbf{R}_{RT}$ is different. $\mathbf{R}_{RT}$ indicates a gap between the number of followers and the popularity of one's tweets and brings a new perspective in influence in Twitter.

5. TRENDING THE TRENDS

In Section 3 we have looked at the topological characteristics of the Twitter network and learned of low reciprocity in Twitter. If we interpret the act of following as subscribing to tweets, then Twitter serves more as an information spreading medium than an online social networking service. Then what information does spread on Twitter? In this section we examine what topics become trending topics and how trending topics rise in popularity, spread through the followers' network, and eventually die.

As described in Section 2.1, we obtain $4,266$ unique trending topics from June 3rd to September 25th, 2009. This period includes big events such as Apple's Worldwide Developers Conference, the E3 Expo, NBA Finals, and the Miss Universe Pageant; tragic events of Michael Jackson's death and the Air France Flight 447 plunge; the Iran election; theatre release of Harry Potter and the Half-Blood Prince; global product releases of iPhone 3GS, Snow Leopard, Zune HD, etc. There are also some hashtags (e.g., #whateverhappened and #thingsihate) that represent Twitter-only trends.

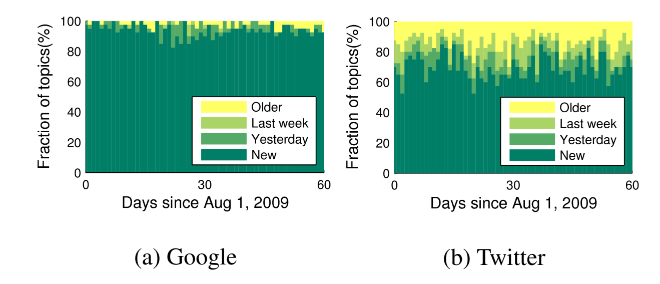
5.1 Comparison with Trends in Other Media

To answer what topics are popular in Twitter, we compare Twitter's trending topics with those in other media, namely, Google Trend and CNN headlines. Google search is the most popular service people use to search for information in today's Internet. The search keywords represent topics users are interested in and popular keywords represent hot trends, although the detailed mechanism of Google Trend is unknown. Search keywords have become a good indicator to understand activities in the real world [9].

We have collected top 40 search keywords per day from Google Trend during the same period as our Twitter data collection. We have also extracted top 40 trending topics per day on Twitter. We first compare the Google keywords to the trending topics in Twitter. We consider a search keyword and a trending topic a match if the length of the longest common substring is more than 70% of either string. Only 126 (3.6%) out of $3,479$ unique trending topics from Twitter exist in $4,597$ unique hot keywords from Google. Most of them are real world events, celebrities, and movies (e.g., mlb draft, tsunami, michael jackson, and terminator)

We also compare the freshness of topics in Google Trend and Twitter trending topics. In Figure 9 we plot how many topics are fresh, a day old, a week old, or longer. On average 95% of topics each day are new in Google while only 72% of topics are new in Twitter. Interactions among users, e.g., retweet, reply, and mention, are prevalent in Twitter unlike Google search, and such interactions might be a factor to keep trending topics persist.

How close are trending topics to CNN Headline News in time and coverage? We collected CNN Headline News of our Twitter data collection period and conducted preliminary analysis. From a subset of trending topics that we have matched against CNN Headline News more than half the time CNN was ahead in reporting. However, some news broke out on Twitter before CNN and they are of live broadcasting nature (e.g., sports matches and accidents). Our preliminary results confirms the role of Twitter as a media for breaking news in a manner close to omnipresent CCTV for collective intelligence.



(a) Google (b) Twitter

Figure 9: The age of the trending topics from Google and Twitter

5.2 Singleton, Reply, Mention, and Retweet

A tweet can be just a statement made by a user, or could be a reply to another tweet. Or a retweet, which refers to a common practice in Twitter to copy someone else's tweet as one's own, sometimes with additional comments. Retweets are marked with either "RT" followed by '@user id' or "via @user id". Retweet is considered the feature that has made Twitter a new medium of information dissemination. People often write a tweet addressing a specific user. We call such a tweet a mention. Both replies and mentions include '@' followed by the addressed user's Twitter id. If a tweet has no reply or a retweet, then we call it a singleton.

Trend	# of users	Singleton	Reply	Mention	RT
rt &	184,351	0.046	0.042	0.070	0.842
#iranelection	120,320	0.559	0.071	0.062	0.307
tehran	69,346	0.661	0.049	0.060	0.230
joe wilson	82,685	0.636	0.118	0.049	0.197
van jones	51,883	0.646	0.129	0.048	0.176
bill clinton	67,024	0.718	0.072	0.039	0.170
iran	292,150	0.712	0.080	0.039	0.169
north korea	61,528	0.707	0.103	0.038	0.152
president obama	172,807	0.709	0.098	0.042	0.151
obama	403,534	0.709	0.101	0.040	0.150
senator ted kennedy	96,201	0.773	0.043	0.043	0.141
oprah	69,596	0.592	0.182	0.085	0.141
kayne west	214,052	0.743	0.069	0.047	0.140
marvel	68,120	0.690	0.118	0.052	0.139
jackass	59,496	0.720	0.099	0.043	0.138
ted kennedy	82,844	0.754	0.067	0.046	0.134
president	165,304	0.713	0.109	0.046	0.133
kanye	300,222	0.690	0.128	0.049	0.133
remembering 9	53,940	0.707	0.084	0.081	0.128
michael vick	93,790	0.662	0.172	0.042	0.125

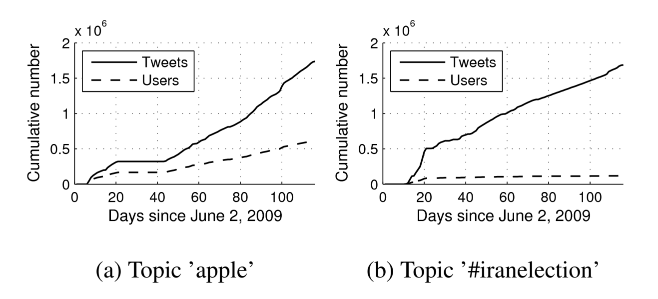
Figure 10: Topics ranked by RT proportion (# of users > 50,000)

Among all tweets mentioning $4,266$ unique trending topics, singletons are most common, followed by replies and retweets. Mentions are least common in tweets. However, the proportions of singletons, replies, mentions, and retweets vary greatly depending on

the topic. In Figure 10 we list the top 20 topics ranked by the proportion of retweets. All but two topics are about offline news, and the remaining two are about a campaign ('remembering 9') and, we suspect, a bug ('rt &') of Twitter in extracting frequent words from retweets.

5.3 User Participation in Trending Topics

How many topics does a user participate on average? Out of 41 million Twitter users, a large number of users (8, 262, 545) participated in trending topics and about 15% of those users participated in more than 10 topics during four months.

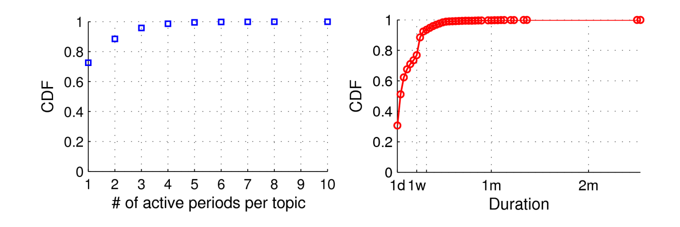


(a) Topic 'apple' (b) Topic '#iranelection'

Figure 11: Cumulative numbers of tweets and users over time

Long-lasting topics with an increasing number of tweets do not always bring in new users into the discussion. In Figure 11 the two topics 'apple' and '#iranelection' have similar numbers of tweets, but the number of user participating in 'apple' is five times larger than that of '#iranelection'. Moreover, the pace at which new users write on the topic '#iranelection' slows down after the first 20 days. We find that there exist core members generating many tweets over a long time period for that particular trending topic.

5.4 Active Period of Trends



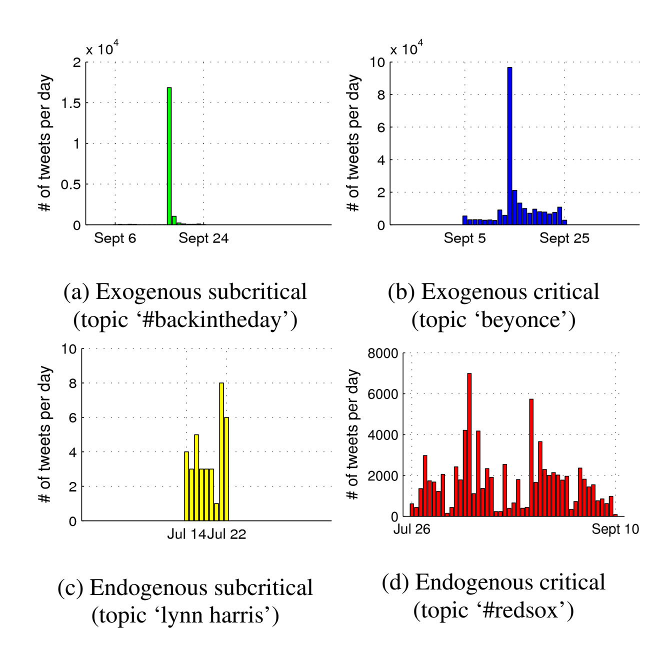
(a) # of active periods / topic (b) Duration of active period

Figure 12: Cumulative fraction

A trending topic does not last forever nor dies to never come back. If we consider a trending topic inactive if there is no tweet on the topic for 24 hours, then we have 6, 058 active periods from 4, 266 trending topics. In Figure 12 we plot the CDF of the active periods and find that 73% topics have a single active period. About 15% of topics have 2 active periods and 5% have 3. Very few have more than 3 active periods.

Most of the active periods are a week or shorter. In Figure 12 we see that 31% of periods are 1 day long, and only 7% of periods are longer than 10 days. There are, however, a few long-lasted topics that have been active for more than two months. The longest lasted for 76 days, and the corresponding topic was 'big brother.'

How many tweets does a topic attract at the beginning, in the middle and near the end of the topic duration? Crane and Sornette present a model that categorizes the response function in a social system [7]. Their model takes into consideration whether the factor behind an event is endogenous or exogenous and whether a user can spread the news about the event to others or not (critical or subcritical). They evaluate their model using 5 million videos of YouTube and label videos as viral, quality, and junk solely based on the quantitative analysis of the number of views and time. Just as on YouTube, there are endogenous and exogenous factors that push a topic to the top trending topic list and the spread of the topic follows an epidemic cascade through the network of followers. We apply their classification methodology on the number of tweets and their times, and classify trending topic periods into the following four categories: exogenous subcritical, exogenous critical, endogenous subcritical, and endogenous subcritical. Sample topics from each category are shown in Figure 13. We confirm that each category has its unique popularity pattern.



(a) Exogenous subcritical (topic '#backintheday') (b) Exogenous critical (topic 'beyonce')

(c) Endogenous subcritical (topic 'lynn harris') (d) Endogenous critical (topic '#redsox')

Figure 13: The examples of classified popularity patterns

Manual inspection of the topics that fall into the exogenous critical class reveal that they are mostly timely breaking news, which we refer as headline news. The topics in the endogenous critical class are of more lasting nature: professional sports teams, cities, and brands. We label them as persistent news. Those exogenous subcritical topics have hashtags, such as #thoughtsintheclub and #thingsihate, catching a limited subset of users' attention and eventually dying out. We call them ephemeral.

	Subcritical		Critical	
Exo.	31.5%	(1,905)	54.3%	(3,290)
Endo.	6.9%	(419)	7.3%	(444)

Table 1: # of topics in each category

The numbers and percentage of active periods in each class are shown in Table 1. The largest number falls into the exogenous critical class. We claim that Twitter users tend to talk about topics from headline news and respond to fresh news.

6. IMPACT OF RETWEET

We have seen how trending topics rise in popularity and eventually die in Section 5. Then how exactly does information spread on Twitter? Retweet is an effective means to relay the information beyond adjacent neighbors. We dig into the retweet trees constructed per trending topic and examine key factors that impact the eventual spread of information.

6.1 Audience Size of Retweet

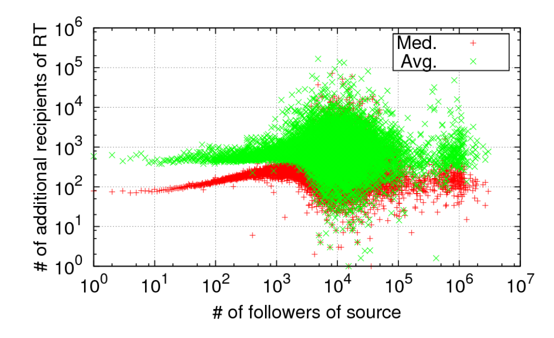


Figure 14: Average and median numbers of additional recipients of the tweet via retweeting

People subscribe to mass media in various forms: radio, TV, and newspapers. They are immediate recipients and consumers of the news the established media produce. On Twitter people acquire information not always directly from those they follow, but often via retweets. Assuming a tweet posted by a user is viewed and consumed by all of the user's followers, we count the number of additional recipients who are not immediate followers of the original tweet owner. Figure 14 displays its average and median per tweet against the number of followers of the original tweet user. The median lies almost always below the average, indicating that many tweets have a very large number of additional recipients. Up to about $1,000$ followers, the average number of additional recipients is not affected by the number of followers of the tweet source. That is, no matter how many followers a user has, the tweet is likely to reach a certain number of audience, once the user's tweet starts spreading via retweets. This illustrates the power of retweeting. That is, the mechanism of retweet has given every user the power to spread information broadly. We recall that influentials by the number of retweets are dissimilar with those by the number of followers or PageRank. Individual users have the power to dictate which information is important and should spread by the form of retweet, which collectively determines the importance of the original tweet. In a way we are witnessing the emergence of collective intelligence.

6.2 Retweet Trees

Knowing that retweet actually delivers information to far more people than a source's immediate followers, we are now interested in how far and deep retweets travel in Twitter. In order to answer the question we build an information diffusion tree of every tweet that is retweeted and call it a retweet tree. All retweet trees are subgraphs of the Twitter network.

We illustrate all the retweet trees of the topic 'air france flight' in Figure 15. In every connected component different colors represent different tweets. The forest of retweet trees has a large number of one or two-hop chains. We find interesting retweet patterns such as repetitive retweet and cross-retweet; the former is repeatedly

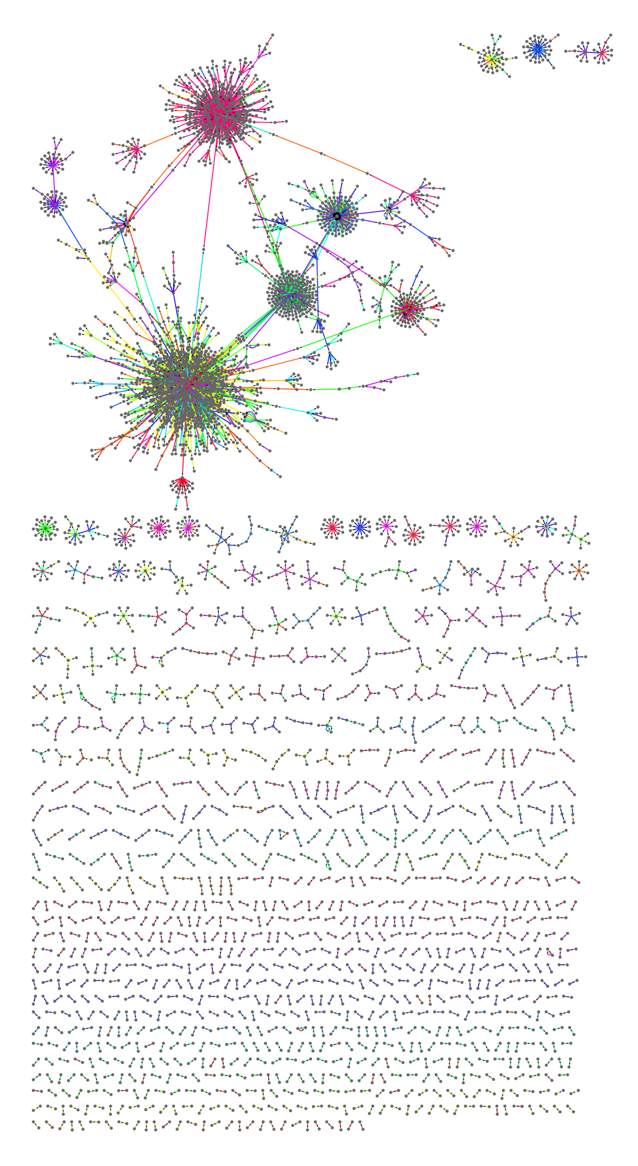


Figure 15: Retweet trees of 'air france flight' tweets

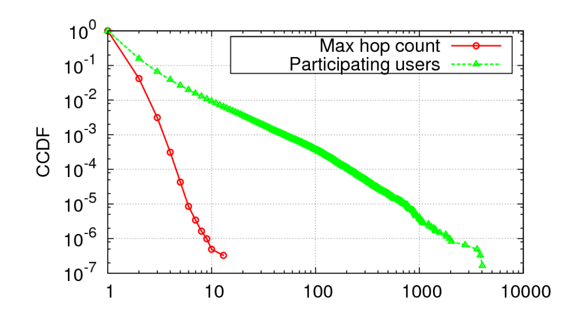


Figure 16: Height and participating users in retweet trees

retweeting the same tweet, and cross-retweet is retweeting each other.

In Figure 16 we plot the CCDFs of the retweet tree heights and the number of users in a retweet tree. The height of 1 is the most

common claiming 95.8%. As 97.6% of node pairs have less than 6 degrees of separation, all retweet trees but for a handful have a height smaller than 6, and no tree goes beyond 11 hops. The distribution of the users in a retweet tree follows power-law. This retweet tree analysis demonstrates how retweets spread and how many get involved.

6.3 Temporal Analysis of Retweet

We have seen in Section 6.2 that most retweet trees have a height of one, but retweets reach a good number of people no matter how many followers the tweet source has. Here we investigate how soon retweets appear and how long they last. Figure 17 plots the time lag from a tweet to its retweet. Half of retweeting occurs within an hour, and 75% under a day. However about 10% of retweets take place a month later,

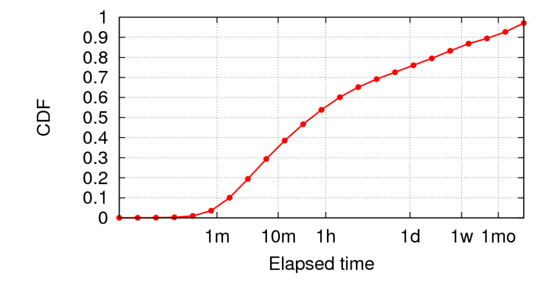


Figure 17: Time lag between a retweet and the original tweet

In Figure 18 we plot the time lag between two nodes on a retweet tree. As most retweet trees are one-hop deep, the time lag on the first hop is spread out, with the median at just under 1 hour and the inter-quartile range expanding from a few minutes to more than a day. What is interesting is from the second hop and on is that the retweets two hops or more away from the source are much more responsive and basically occur back to back up to 5 hops away. Cha *et al.* reports that favorite photos diffuse in the order of days in Flickr [4]. The strength of Twitter as a medium for information diffusion stands out by the speed of retweets.

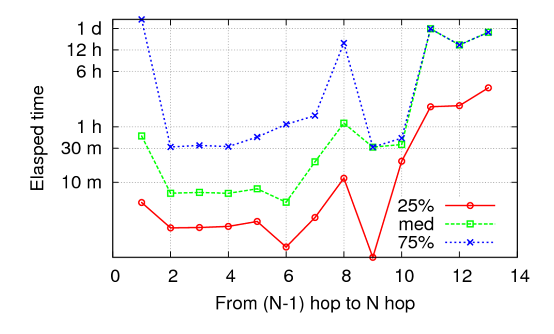


Figure 18: Elapsed time of retweet from $(n-1)$ hop to n hop

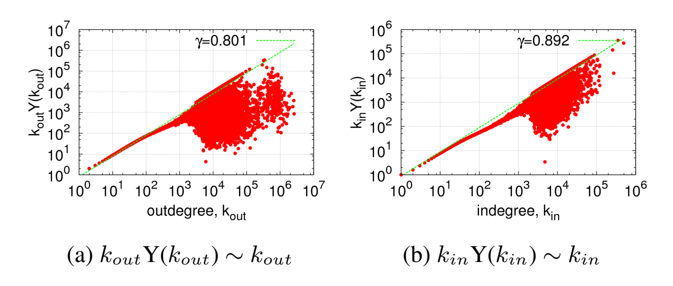
6.4 Favoritism in Retweet

When a user retweets, the user may or may not retweet evenly from those whom the user follows. Also from the perspective of a user who gets retweeted, the retweet may or may not take place evenly among one's followers. How even is the information diffusion in retweet? To answer this question we investigate disparity [2] in retweet trees.

For each user i we define $|r_{ij}|$ as the number of retweets from user j. The $Y(k,i)$ is defined as follows:

$$Y(k,i) = \sum_{j=1}^{k} \left\{ \frac{|r_{ij}|}{\sum_{l=1}^{k} |r_{il}|} \right\}^2 \qquad (3)$$

$Y(k)$ represents $Y(k,i)$ averaged over all nodes that have k outgoing (incoming) edges. Here an edge represents a retweet. When retweeting occurs evenly among followers, then $kY(k) \sim 1$. If most of retweeting occurs within a subset of followers, then $kY(k) \sim k$. For outgoing links, similar interpretation applies. Both Figures 19(a) and 19(b) shows a linear correlation up to $1,000$ followers. The linear correlation to k represents favoritism in retweets: people only retweets from a small number of people and only a subset of a user's followers actually retweet. Chun *et al.* also report that favoritism exists in conversation from guestbook logs of Cyworld, the biggest social networks in Korea [5].



(a) $k_{out}\mathrm{Y}(k_{out}) \sim k_{out}$ (b) $k_{in}\mathrm{Y}(k_{in}) \sim k_{in}$

Figure 19: Disparity in retweet trees

7. RELATED WORK

Online social networks and social media

The rising popularity of online social networking services has spurred research into their characteristics and recent work has forayed into characteristics beyond crawled data [3, 39].

Twitter is less than three years old, but has attracted much attention in the past two years. Java *et al.* conduct preliminary analysis of Twitter in 2007 [14]. Their dataset covers about $76,000$ users and $1,000,000$ posts. They find user clusters based on user intention to topics by clique percolation methods. Krishnamurthy *et al.* also analyze the user characteristics by the relationships between the number of followers and that of followings [17]. Zhao and Rosson qualitatively investigate the motivation of using Twitter [40]. Huberman *et al.* reports that the number of friends is actually smaller than the number of followers or followings [11]. Jansen conducts preliminary analysis of word-of-mouth branding in Twitter [13]. Our work marks the first to look at the entire Twittersphere.

Information cascades

Information diffusion is a process that a new idea or an action widely spreads through communication channels [32]. This area is extensively researched from sociology, marketing, and epidemiology [15, 19, 30, 33]. The success of online social networks opens a new problem of large-scale information diffusion. Topic propagation in blogspace [10], linking patterns in blog graph [21], favorite photo marking in a social photo sharing service [4], fanning in Facebook [34], Internet chain letter forwarding [24], and meme tracking in news cycles [20] all report on large-scale information diffusion online. We treat retweet trees as communication channels of information diffusion and observe that retweets reach a large audience and spread fast.

8. CONCLUSIONS

We have crawled the entire Twittersphere and obtained 41.7 million user profiles, 1.47 billion social relations, $4,262$ trending topics, and 106 million tweets. In its follower-following topology analysis we have found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks [28]. Among reciprocated users we observe some level of homophily. In order to identify influentials on Twitter, we have ranked users by the number of followers and by PageRank and found two rankings to be similar. If we rank by the number of retweets, then the ranking differs from the previous two rankings, indicating a gap in influence inferred from the number of followers and that from the popularity of one's tweets. Ranking by retweets exposes the influence of other media in a novel perspective. We have analyzed the tweets of top trending topics and reported on the temporal behavior of trending topics and user participation. We then classify the trending topics based on the active period and the tweets and show that the majority (over 85%) of topics are headline or persistent news in nature. A closer look at retweets reveals that any retweeted tweet is to reach an average of $1,000$ users no matter what the number of followers is of the original tweet. Once retweeted, a tweet gets retweeted almost instantly on the 2nd, 3rd, and 4th hops away from the source, signifying fast diffusion of information after the 1st retweet.

Twitter with its open API to crawl, one-sided nature of relationship, and the retweet mechanism to relay information offers an unprecedented opportunity for computer scientists, sociologists, linguists, and physicists to study human behavior. Our work is the first step towards exploring the great potentials of this new platform.

9. ACKNOWLEDGEMENTS

We are grateful to Meeyoung Cha, Yong-Yeol Ahn and Young-Ho Eom for helpful discussions. We also thank anonymous reviewers for their valuable comments and suggestions. This work was supported by the IT R&D program of MKE/KEIT [2008-F-016-02, "CASFI: High-Precision Measurement and Analysis Research"].

10. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proc. of the 16th international conference on World Wide Web*. ACM, 2007.

[2] E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A. L. Barabási. Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839–843, February 2004.

[3] F. Benevenut, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *Proc. of ACM SIGCOMM Internet Measurement Conference*. ACM, 2009.

[4] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the Flickr social network. In *Proc. of the 18th international conference on World Wide Web*. ACM, 2009.

[5] H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of Cyworld. In *Proc. of the 8th ACM SIGCOMM Internet Measurement Conference*. ACM, 2008.

[6] Clean Tweets. http://blvdstatus.com/clean-tweets.html.

[7] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. of the National Academy of Sciences*, 105(41):15649–15653, 2008.

[8] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. In *Proc. of the 14th annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2003.

[9] Flu Trends. http://www.google.org/flutrends/.

[10] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. of the 13th international conference on World Wide Web*. ACM, 2004.

[11] B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. arXiv:0812.1045v1, Dec 2008.

[12] HubSpot. State of the twittersphere. http://bit.ly/sotwitter, June 2009.

[13] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. In *Proc. of the 27th international conference extended abstracts on Human factors in computing systems*. ACM, 2009.

[14] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, 2007.

[15] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003.

[16] M. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[17] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proc. of the 1st workshop on Online social networks*. ACM, 2008.

[18] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.

[19] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *Proc. of the 7th ACM conference on Electronic commerce*. ACM, 2006.

[20] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009.

[21] J. Leskovec and E. Horvitz. Worldwide buzz: Planetary-scale views on an instant-messaging network. Technical report, Microsoft Research, June 2007.

[22] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proc. of the 17th international conference on World Wide Web*. ACM, 2008.

[23] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2005.

[24] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. of the National Academy of Sciences*, 105(12):4633–4638, 2008.

[25] F. McCown and M. L. Nelson. Agreeing to disagree: search engines and their public interfaces. In *Proc. of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2007.

[26] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[27] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[28] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, Sep 2003.

[29] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[30] R. Pastor-Satorras and A. Vespignani. Epidemics and immunization in scale-free networks. arXiv:cond-mat/0205260v1, May 2002.

[31] E. Reinikainen. `#iranelectioncyberwarguideforbeginners`. http://goo.gl/pZvi, June 2009.

[32] E. M. Rogers. *Diffusion of Innovations*. Free Press, 5 edition, August 2003.

[33] D. Strang and S. Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual Review of Sociology*, 24:265–290, 1998.

[34] E. Sun, I. Rosenn, C. Marlow, and T. Lento. Gesundheit! modeling contagion through facebook news feed. In *Proc. of International AAAI Conference on Weblogs and Social Media*, 2009.

[35] The New York Times. http://bits.blogs.nytimes.com/2009/07/07/spammers-shorten-their-urls/.

[36] Twitter Search API. http://apiwiki.twitter.com/Twitter-API-Documentation.

[37] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, Jun 1998.

[38] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. of the third ACM international conference on Web search and data mining*. ACM, 2010.

[39] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proc. of the 4th ACM European conference on Computer systems*. ACM, 2009.

[40] D. Zhao and M. B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 2009.